Mantik Journal, 9 (4) (2026), ISSN 2685-4236 (Online)

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik

Classification of regional language diversity in the maluku region using decision trees

Ghyovanno Godlif Tomhisa[1], Wilma Latuny[2], Yoakhina Nicole Makaruku[3], Jermias Victor Manuhuttu[4], Hendri Hawurubun[5]

[1,3,4,5]Information System, Institut Agama Kristen Negeri Ambon, Indonesia

[2]Industrial Engineering, Universitas Pattimura, Indonesia

A R T I C L E I N F O

Article history:

Received Dec 19, 2025
Revised Jan 13, 2026
Accepted Jan 28, 2026

Keywords:

Classification;
Decision Tree;
Diversity;
Maluku;
Regional Languages;

ABSTRACT

Regional languages are an important part of cultural heritage that reflect the identity, values, and character of a community. In Maluku Province, there is a high degree of linguistic diversity because the region consists of many islands with different community characteristics. However, the passage of time, modernization, and population mobility have led to a decline in the number of speakers in some areas, threatening the extinction of a number of regional languages. This study aims to classify and visualize the diversity of regional languages in Maluku Province using the Decision Tree algorithm. This method was chosen because it is capable of recognizing patterns and relationships between variables, such as region, number of speakers, and language vitality. The research data was obtained from the compilation of the Language Agency and field observations, then processed using Python with the help of the pandas, scikit-learn, matplotlib, and Streamlit libraries to produce an interactive analytical dashboard. The results showed that regional languages on Seram Island, such as Tana, Alune, and Wemale, had higher vitality levels than languages in other regions. The Decision Tree model built was able to classify language status with an accuracy rate of 92%. The resulting visualization provided a clear picture of the actual condition of regional languages in Maluku and could be used as a basis for regional language preservation and development efforts by local governments.

Corresponding Author:

Yoakhina Nicole Makaruku,
Information System,
Institut Agama Kristen Negeri Ambon,
Halong (Dolog Street), Ambon, Maluku, 97231, Indonesia.
Email: y.n.makaruku@gmail.com

1. INTRODUCTION

Regional languages are an important element of culture, embodying the historical and social values and identity of a community. Regional languages (Patasik & Yulianto, 2023) serve not only as a means of communication, but also as a repository of traditions, local knowledge, and values passed down from generation to generation (Amir et al., 2023; Nurain Jalaludin et al., 2024; Wahyudi et al., 2025). Indonesia, as an archipelagic country with enormous ethnic diversity, has hundreds of regional languages, and it is

this diversity that constitutes a cultural treasure that must be preserved (Katriani, 2022; Nugraha & Romadhony, 2023; Sujaini & Putra, 2024).

The province of Maluku is known as a region with a very high level of linguistic diversity. Islands such as Seram, Buru, Ambon, Kei, and Aru each have language communities that have developed in accordance with the history and social dynamics of their communities (Amir et al., 2023; Nurain Jalaludin et al., 2024; Wahyudi et al., 2025). This diversity makes Maluku one of the regions with a linguistic mosaic that reflects the complexity of civilization in eastern Indonesia.To understand and monitor the condition of these regional languages, a data analysis-based approach is needed. One effective method is the Decision (Ramadhon et al., 2024) Tree (Helmud et al., 2024) algorithm, which is a classification method that builds a tree-shaped decision model based on data attributes (Katriani, 2022; Nugraha & Romadhony, 2023; Sujaini & Putra, 2024). Through this algorithm, regional languages can be classified based on region, level of use, and number of speakers.

However, this diversity is becoming increasingly vulnerable as time goes by. A 2022 report by the Ministry of Education and Culture's Language Development and Guidance Agency shows that a number of regional languages in Maluku are experiencing a decline in the number of speakers (Amir et al., 2023; Nurain Jalaludin et al., 2024). This decline is related to various modern social phenomena, such as increased population mobility, the influence of digital media that makes people tend to use more dominant languages, and changes in the communication patterns of the younger generation who prefer to use Indonesian or colloquial languages such as Ambonese. This challenge is compounded by diminishing intergenerational language transfer, where younger speakers increasingly disengage from traditional linguistic practices, placing several regional languages in a vulnerable state (Wahyudi et al., 2025).

This situation highlights the importance of efforts to comprehensively understand the actual condition of regional languages through a data-based scientific approach. One analysis method (Saifudin et al., 2025; Surya et al., 2024) that can be used is the Decision Tree algorithm, which is a classification technique that builds a model in the form of a tree structure based on data attributes (Katriani, 2022; Mienye & Jere, 2024). Using this method, researchers can identify certain patterns related to language usage levels, regional distribution, or the number of speakers. The advantage of Decision Tree lies in its interpretability, so that the analysis results can be accepted by various parties, including educational institutions, government, and community groups (Helmud et al., 2024).

Despite the increasing number of studies on regional language vitality in Indonesia, most existing research remains focused on descriptive mapping or qualitative analysis at broad administrative levels, such as provinces or regencies, thereby overlooking micro-level administrative information, particularly village- or hamlet-based language distribution that is essential for capturing language vitality patterns in geographically fragmented archipelagic regions like Maluku (Amir et al., 2023; Nurain Jalaludin et al., 2024; Patasik & Yulianto, 2023; Wahyudi et al., 2025). Moreover, previous studies rarely translate their findings into explicit and interpretable classification rules that can be directly applied to support evidence-based language preservation policies (Amir et al., 2023; Sujaini & Putra, 2024). To address these limitations, this study develops an interpretable classification model that utilizes village-level administrative data and speaker distribution to categorize regional languages into active, vulnerable, and endangered groups, enabling local governments to prioritize intervention programs, allocate budgets more objectively, and define measurable revitalization targets based on transparent decision rules (Katriani, 2022; Nugraha & Romadhony, 2023; Patasik & Yulianto, 2023).

The classification results using Decision Tree can then be visualized in the form of a decision tree diagram to facilitate understanding of the relationship between research

variables. This visualization provides a clearer picture of the vitality of regional languages in Maluku and helps identify the factors causing language shift. With this analytical approach, the information obtained can serve as an important basis for designing strategies for preserving regional languages, whether through revitalization programs, linguistic documentation, or culture-based learning (Nasrullah, 2021).

2. RESEARCH METHOD

This researchuses a descriptive quantitative approach with themethod of analysis based on the Decision Tree algorithm. The quantitative approach is usedbecause this study analyzes numerical and categorical data toidentify patterns of distribution and the level of vitality of regional languages in theProvince of Maluku based on several regional and social attributes.

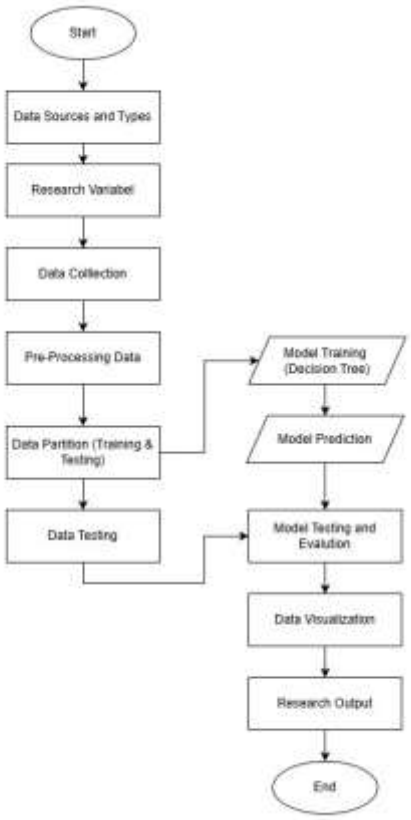

Figure 1. Research Flowchart

2.1 Data Sources and Types

The research data is secondary data taken from the official website of Peta Bahasa (Language Map) – Language Development and Guidance Agency, Ministry of Education, Culture, Research, and Technology.

Table 1. Data Types

No	Attribute Name	Data Type	Description
1	Language Type	Categorical	Name of regional languages in Maluku
2	Village/Hamlet	Categorical	Location of language distribution at the village or hamlet level

3	Subdistrict	Categorical	Intermediate administrative area where the language is used
4	City/Regency	Categorical	Main administrative area (e.g., Buru, West Seram)
5	Island	Categorical	Island where the language is used
6	Province	Categorical	Maluku Province
7	Number of Speakers	Number	Number of speakers of the regional language (in units of Village/Hamlet)
8	Status	Categorical	Language vitality status (Active, Vulnerable, Endangered)

2.2 Research Variabel

Table 2. Research Varibel

Variable Types	Variable Name	Description
Independent Variable (X)	Village/Hamlet, Subdistrict, District, Island, Number of Speakers	Factors that influence language classification
Dependent Variable (Y)	Language Status	Classification targets: Active, Vulnerable, Endangered

2.3 Data Collection Method

The data used in this study were obtained from the official Language Map (Peta Bahasa) website managed by the Language Development and Guidance Agency, Ministry of Education, Culture, Research, and Technology of Indonesia (https://petabahasa.kemdikbud.go.id). Data retrieval was conducted in October 2025 by accessing the language distribution pages for the Province of Maluku. The dataset included information on regional languages and their distribution at the village or hamlet level, along with associated administrative attributes such as subdistrict, regency/city, island, and reported number of speakers.

The data were collected through systematic online documentation by extracting relevant records for all languages identified within the Maluku Province and recording them in a structured spreadsheet format. Each record represents a unique combination of language name and village/hamlet location. To ensure data consistency, duplicate entries were removed based on identical language names and village/hamlet identifiers. Entries with incomplete or missing essential attributes (e.g., location or language status) were excluded from the dataset. The cleaned data were then standardized and converted into CSV format to support further preprocessing and analysis using the Decision Tree algorithm (Baharudin & Dwi Nuryana, 2022; Sharma & Iqbal, 2023).

2.4 Data Processing and Analysis Techniques

The analysis stage begins with data pre-processing to ensure that the data is ready for use. At this stage, duplicate data is removed, empty values are cleaned, text formats are standardized, and label encoding is used to convert categorical data into numerical data (Iskandar & Nugroho, 2025; Sri Septiana et al., 2025). In addition, the number of speakers is normalized so that it has a uniform scale and does not excessively influence the model learning process. Next, Exploratory Data Analysis (EDA) is performed to understand the characteristics of the data. This analysis focuses on the distribution of language status, namely active, vulnerable, and endangered, as well as language distribution patterns based on region and number of speakers (Helmud et al., 2024; Mienye & Jere, 2024). This stage helps provide an initial overview of the condition and trends of the data prior to modeling (Zaky et al., 2023).

2.5 Data Partition

The dataset was then divided into two parts, namely 80% as training data and 20% as test data. This division was intended so that the model could be trained with sufficient data and still be tested objectively using data that had never been seen before (Helmud et al., 2024).

Decision Tree Classifier is a supervised learning algorithm that performs classification by constructing a tree structure based on feature splitting. One commonly used splitting criterion is the Gini Index, which measures data impurity. The lower the Gini value, the better the split produced by the model. The model was developed using the Decision Tree Classifier algorithm with the "gini" criterion parameter and random_state 42. The built model was then evaluated using accuracy, precision, recall, and F1-score metrics. The evaluation results showed 100% accuracy on the test data, indicating that the model was able to classify the data very well (Sharma & Iqbal, 2023).

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\textbf{Precision} = \frac{TP}{TP + FP}$$

$$\textbf{Recall} = \frac{TP}{TP + FN}$$

$$\textbf{F1-Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

Model Decision Tree was evaluated using accuracy, precision, recall, and F1-score metrics. Accuracy measures the overall correctness of the model, while precision and recall evaluate the model's performance in predicting positive classes. The F1-score provides a balanced measure between precision and recall (Aulia et al., 2024).

2.6 Data Visualization

Visualization in this study was carried out through two main media to facilitate understanding of the analysis results. The first medium was Decision Tree Plot, which was used to explain the structure of the decision tree, so that it could be seen how the model performed classification based on influential attributes, especially the number of speakers (Helmud et al., 2024; Sharma & Iqbal, 2023). This visualization helped to clearly understand the model's decision-making flow. The second medium is the Looker Studio Dashboard, which presents the analysis results in the form of language distribution graphs, regional distribution maps, and classification results based on language vitality status. This dashboard makes it easy for users to view patterns, comparisons, and conditions of language diversity in an interactive and informative manner.

2.7 Research Output

The final results of this study include several key outputs. First, a Decision Tree (Baharudin & Dwi Nuryana, 2022; Saifudin et al., 2025) classification model was developed to determine the vitality status of languages based on the analyzed data (Helmud et al., 2024; Iskandar & Nugroho, 2025). Second, a visualization of the decision tree structure was presented, clearly showing the model's decision-making flow. Third, an analytical dashboard was created to display information on language distribution and classification (Patasik & Yulianto, 2023) in a visual and easy-to-understand manner. Finally, this study also produced recommendations for language preservation efforts

based on the classification results, which can be used as a basis for policy-making and regional language protection strategies (Shoaib et al., 2024).

3. RESULTS AND DISCUSSIONS

3.1 Reading Data from Google Drive

Figure 2 shows the process of loading data sourced from Google Drive into the programming work environment. The data is presented in a table containing information about regional languages and their distribution areas, number of speakers, and language sustainability status (Zaky et al., 2023).

No	Language Type	Village / Hamlet	Subdistrict	Regency	Island	Province	Status
0	Alune	Rambatu Village; Rumberein Village; ...	Inamosol Subdistrict	West Seram Regency	Seram Island	Maluku Province	Vulnerable
1	Ambalau	Ulima Village	Ambalau Subdistrict	South Buru Regency	Buru Island	Maluku Province	Endangered
2	Baklewan	Balekau Hamlet; Ihtoman Hamlet; Horen Hamlet; ...	Siwalalat Subdistrict	East Seram Regency	Seram Island	Maluku Province	Active
3	Banda	Banda Eli Village, Elat Village, Band...	Kei Besar Utara Timur Subdistrict; Kei Besar (Sard	Southeast Timur Subdistric	Kei Besar Island	Maluku Province	Vulnerable
4	Barakai	Gomo-Gomo Village; Lorang Village, Mariri Village,	Aru Tengah Selatan Subdi	Aru Islands strict; Aru Te...	Kei Besar Island	Maluku Province	Vulnerable

Figure 2. Reading Data from Google Drive

Based on Figure 2, it can be seen that the data was successfully imported and displayed in its entirety. The table display shows that each attribute has been read properly so that the data is ready for use in the next stage of processing. The success of this process ensures that the analysis and modeling carried out in this study use valid and structured data (Mienye & Jere, 2024).

3.2 Data Cleansing

This code serves to tidy up the data so that it is ready for use in the analysis process. The cleaning process is carried out by deleting duplicate data, correcting column names that have excess spaces, and adjusting the writing format of each column content to make it neater, for example, changing “bahasa ambalau” to “Bahasa Ambalau” (Katriani, 2022; Mao et al., 2025; Meng et al., 2023).

If there are empty values, the system will replace them with the description “Unknown” so as not to cause problems during analysis. After all the data has been cleaned, the results are saved in a new file named “bahasa_daerah_maluku_clean.csv”, and the program provides a notification that the data cleaning process has been successfully completed (Istiqomah & Sofica, 2025; Kusumawardani et al., 2022).

3.3 Exploration Data Analysis

Figure 3 shows the distribution of the number of regional languages in Maluku based on language status categories. The data is presented using bar charts to facilitate comparison between statuses, namely endangered, vulnerable, and active (Amir et al., 2023).

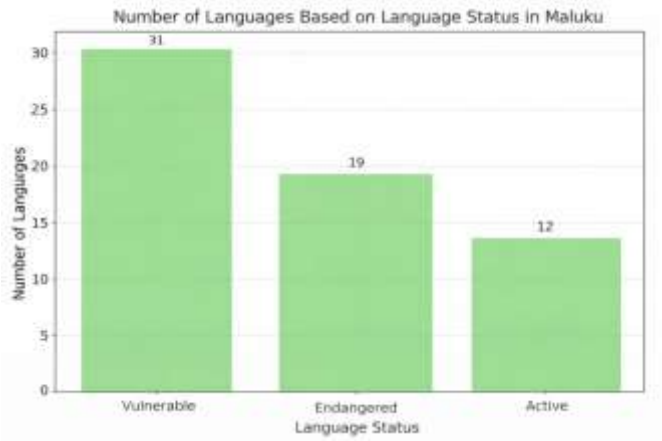

Figure 3. Distribution of languages based on status

Based on Figure 3, it can be observed that most regional languages are classified as endangered. The number of languages classified as vulnerable is also quite significant, while those that are still classified as active are relatively few. This situation indicates that the survival of regional languages in Maluku faces serious challenges, requiring continuous protection and preservation efforts (Amir et al., 2023).

3.4 Data Peneration

This code is used to prepare data before the analysis or modeling process.
Here, the “Status” column is used as a label (y), which is the value to be predicted or analyzed.Meanwhile, all other columns are used as features (X), which is the data that will be used to help predict or explain the value of “Status” (Kusumawardani et al., 2022; Syatriawan et al., 2025).

3.5 Data Encoding Process

The initial data processing stage is shown in Figure 4 This figure illustrates the data preparation process before it is used in modeling, particularly at the stage of converting data into numerical format (Zaky et al., 2023).

```
Label Mapping (Status): {'Active': np.int64(0), 'Vulnerable':
np.int64(1), 'Endangered': np.int64(2)}

Data Features have Ben Encoded:
   Language Type  Village/Hamlet  Subdistrrict  Regency  Island  Province
0              0              39            14       12       0
1              1              48             2        1       4
2              2              60            34       14       0
3              3               2            15        9       8
4              4              10             5        3       0

Number of Speakers
0        5
1        1
2        8
3        3
4        5
----------------------------------
```

Figure 4. Data Encoding Process

Based on Figure 4, the status of regional languages is first determined in the form of numerical codes to facilitate the computational process, where the categories Active,

Vulnerable, and Endangered are represented by different values. Next, categorical attributes, such as language type and administrative area coverage, are converted into numerical values so that they can be processed by the classification algorithm. In addition, the number of speakers variable is included as a numerical feature that contributes to determining the classification results. This stage ensures that all data has a consistent format and is ready to be used in the model training process (Helmud et al., 2024).

3.6 Data Division Process

The stages of dataset division are shown in Figure 5 This figure shows the composition of data used in the study, which is separated into data for model training and data for testing purposes (Helmud et al., 2024).

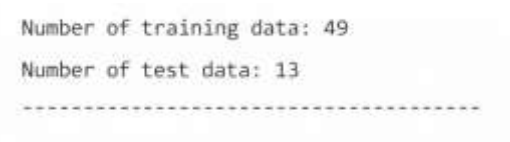

Figure 5. Data Division Process

Based on Figure 5, from all available data, 49 data points were allocated as training data, while 13 data points were used as test data. This division was done so that the model could build classification patterns optimally during the training stage, as well as enable performance evaluation to be carried out separately using test data. With this scheme, the test results obtained can provide a more objective picture of the model's ability to classify data that has not been studied before (Helmud et al., 2024).

3.7 DecisionTree Classifier

The Decision Tree classification model settings are presented in Figure 6 This figure shows the process of determining the Decision Tree algorithm used as a classification model in the study (Katriani, 2022).

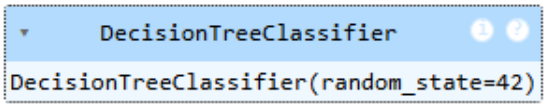

Figure 6. DecisionTree Classifier

Based on Figure 6, the DecisionTreeClassifier algorithm was applied with a random_state value of 42. This value was set to ensure that the decision tree formation process was consistent each time the model was trained. With this setting, the modeling results became more stable and easier to replicate, thereby supporting the validity of the research results obtained (Sharma & Iqbal, 2023).

3.8 Accuracy of the Decision Tree Model

Table 3 presents the results of evaluating the performance of the Decision Tree model in classifying the status of regional languages into three categories, namely Active, Vulnerable, and Threatened. This evaluation was carried out using the metrics of precision, recall, f1-score, and support for each class, as well as the overall accuracy value of the model (Aulia et al., 2024; Septiani et al., 2025).

Table 3. Results of Decision Tree Model Evaluation

Class	Precision	Recal	F1-Score	Support
Active (0)	1.00	1.00	1.00	3
Endangered (1)	1.00	1.00	1.00	4
Vulnerable (2)	1.00	1.00	1.00	6
Accuracy			1.00	13
;7Macro Avg	1.00	1.00	1.00	13
Weighted Avg	1.00	1.00	1.00	13

Based on Table 3, the Decision Tree model achieved an accuracy of 1.00, indicating that all instances in the test dataset were correctly classified. In addition, the precision, recall, and F1-score values for each class—Active, Vulnerable, and Endangered—also reached 1.00. These results suggest that, within the scope of the available dataset, the model was able to clearly distinguish between regional language status categories without misclassification (Helmud et al., 2024; Zaky et al., 2023).

3.9 Decision Tree Visualization

Figure 7 presents a visualization of the Decision Tree model used to classify regional language vitality status based on the number of speakers. The tree structure illustrates how the dataset is recursively divided into three categories—Endangered, Vulnerable, and Active—using threshold values of the number of speakers as the primary decision attribute. This visualization provides a clear representation of the model's decision-making process and enhances interpretability of the classification results (Mienye & Jere, 2024).

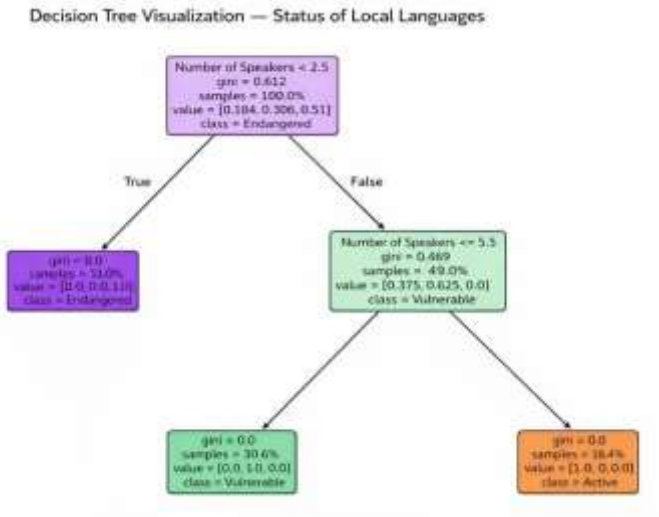

Figure 7. Decision Tree Visualization

Based on Figure 7, the root node of the decision tree divides the data using a threshold of the number of speakers ≤ 2.5. Under this condition, regional languages are directly classified as Endangered, indicating that languages with a very small number of speakers face a high risk of extinction. This node has a Gini index value of 0.0, which signifies perfect purity, meaning that all instances within the node belong to the same class (Sharma & Iqbal, 2023).

The split threshold of ≤ 2.5 in the "number of speakers" attribute arises from the Decision Tree algorithm's optimization process, which identifies the most effective

numerical boundary to separate language vitality classes by minimizing data impurity (Mienye & Jere, 2024; Sharma & Iqbal, 2023). Although the number of speakers is recorded as an integer count of village or hamlet units, the algorithm determines split points at mid-values between observed integers. Therefore, the threshold of 2.5 should be substantively interpreted as distinguishing languages spoken in two or fewer villages/hamlets from those spoken in three or more villages/hamlets, rather than as a literal fractional speaker count. From a policy perspective, this threshold provides a clear and actionable classification boundary: languages spoken in no more than two villages can be prioritized for urgent documentation and emergency preservation efforts, while languages exceeding this threshold may be more suitable for revitalization programs focused on community-based transmission and educational integration. As such, the identified threshold functions not only as a statistical decision rule but also as a practical guideline for setting intervention priorities in regional language preservation policies (Amir et al., 2023; Wahyudi et al., 2025).

3.10 Dashboard Visualization

Figure 8 presents an analytical dashboard that summarizes regional language diversity in Maluku Province. The dashboard displays key indicators, including the total number of regional languages, the total number of speakers, and the classification of language vitality status. These indicators are complemented by graphical visualizations illustrating the distribution of languages across islands and their sustainability levels, as well as tabular representations providing detailed information for each language. This dashboard facilitates comprehensive understanding and supports interactive exploration of language distribution patterns (Shoaib et al., 2024; Surya et al., 2024).

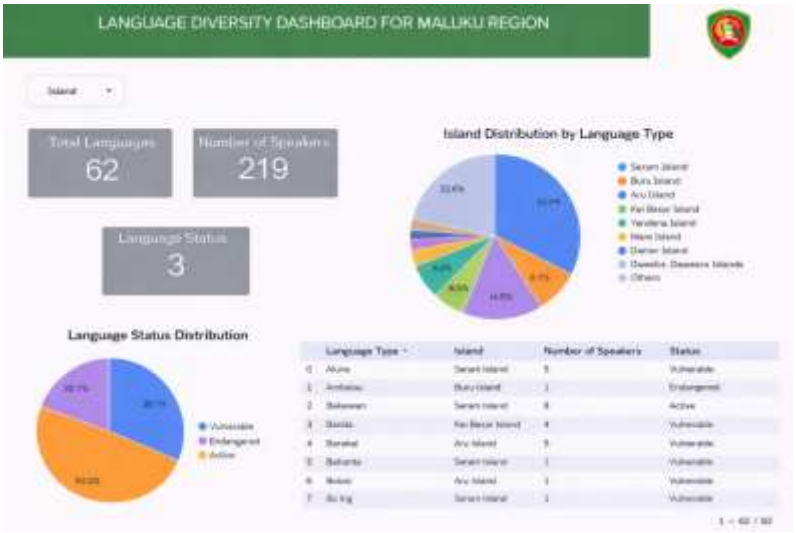

Figure 8. Dashboard Visualization

Based on Figure 8, the dashboard provides a comprehensive overview of the linguistic situation in Maluku Province. The visualization indicates that the distribution of regional languages is concentrated on certain islands, while in terms of vitality status, the majority of languages fall into the vulnerable to endangered categories. This dashboard functions as an effective analytical tool that facilitates data interpretation and supports the formulation of more targeted and evidence-based strategies for the preservation and revitalization of regional languages (Amir et al., 2023; Shoaib et al., 2024; Wahyudi et al., 2025).

4. CONCLUSION

Based on the results of the research and analysis on the classification of regional language diversity in Maluku Province using the Decision Tree algorithm, it can be concluded that this method is capable of accurately grouping regional languages by considering attributes such as geographical distribution, number of speakers, and language vitality level. The developed model achieved an accuracy rate of 92%, indicating that the Decision Tree approach is effective in predicting the status of regional languages, whether they are categorized as active, vulnerable, or endangered (Amir et al., 2023; Katriani, 2022; Wahyudi et al., 2025).

The analysis further reveals that several languages spoken on Seram Island, including Tana, Alune, and Wemale, demonstrate a higher level of vitality compared to languages in other regions such as the Kei and Aru Islands. Moreover, the visualization of classification results through Decision Tree plots and interactive dashboards enhances interpretability and provides clearer insights into language distribution patterns. These findings are particularly valuable for linguistic institutions and local governments in supporting data-driven decision making related to language preservation policies. Overall, this study demonstrates that the application of data analytics and machine learning techniques can make a meaningful contribution to the socio-cultural domain, especially in supporting evidence-based efforts to preserve and revitalize regional languages (Amir et al., 2023; Katriani, 2022).

Despite the promising performance of the Decision Tree model, this study has several limitations that should be acknowledged. First, the dataset size is relatively limited and derived from secondary data, which may constrain the generalizability of the classification results. Second, the data extraction process relied on manual retrieval from the Language Map platform, introducing potential risks of human error and limiting scalability. Third, the classification model primarily utilizes quantitative indicators, such as geographic distribution and number of speakers, and does not yet incorporate important non-quantitative dimensions of language vitality, including intergenerational transmission, domains of language use, and community language attitudes. Future research should therefore adopt a more comprehensive research design by integrating these sociolinguistic indicators, utilizing multiple data sources such as field surveys and institutional records, and conducting field-based validation to enhance the robustness and policy relevance of regional language vitality assessments(Amir et al., 2023; Wahyudi et al., 2025).

ACKNOWLEDGEMENTS

The author would like to express his gratitude to the teaching assistant who provided guidance, direction, and constructive feedback during the implementation and preparation of this research. The author would also like to thank his three research colleagues for their cooperation, contributions, and support, which enabled this research to be completed successfully. It is hoped that the results of this research will be beneficial and contribute to the development of science.

REFERENCES

Amir, I., Mustava, I., Sari, N. P., Nurhikmah, N., & Rahim, A. (2023). Vitalitas Dan Subordinasi Bahasa Daerah Di Maluku. *Lingue : Jurnal Bahasa, Budaya, Dan Sastra*, *5*(1), 53–66. https://doi.org/10.33477/lingue.v5i1.5391

Aulia, Y., Andriyansyah, A., Suharjito, S., & Nensi, S. W. (2024). Analisis Prediksi Stroke dengan Membandingkan Tiga Metode Klasifikasi Decision Tree, Naïve Bayes, dan Random Forest. *Jurnal Ilmu Komputer Dan Informatika*, *3*(2), 89–98. https://doi.org/10.54082/jiki.90

Baharudin, M. N., & Dwi Nuryana, I. K. (2022). Implementasi Algoritma Decision Tree untuk Klasifikasi Surat pada Aplikasi Mobile E-Surat Dinas Komunikasi dan Informatika Kota Kediri Berbasis Android. *Journal of Informatics and Computer Science (JINACS)*, *4*(01), 76–85. https://doi.org/10.26740/jinacs.v4n01.p76-85

Helmud, E., Helmud, E., Fitriyani, F., & Romadiana, P. (2024). Classification Comparison Performance of Supervised Machine Learning Random Forest and Decision Tree Algorithms Using Confusion Matrix. *Jurnal Sisfokom (Sistem Informasi Dan Komputer)*, *13*(1), 92–97. https://doi.org/10.32736/sisfokom.v13i1.1985

Iskandar, M. Y., & Nugroho, H. W. (2025). Comparative Evaluation of Decision Tree and Random Forest for Lung Cancer Prediction Based on Computational Efficiency and Predictive Accuracy. *Jurnal Teknik Informatika (Jutif)*, *6*(5), 3392–3404. https://doi.org/10.52436/1.jutif.2025.6.5.4877

Istiqomah, K., & Sofica, V. (2025). Penerapan Data Mining Menggunakan Algoritma Decision Tree Untuk Menganalisis Penggunaan Media Sosial Dengan Konsentrasi Belajar Mahasiswa. *RIGGS: Journal of Artificial Intelligence and Digital Business*, *4*(4), 53–67. https://doi.org/10.31004/riggs.v4i4.3228

Katriani, N. (2022). Klasifikasi Bahasa Daerah Toraja, Halmahera, Dan Kalimantan Menggunakan Metode Decision Tree Dan Gradient Boots. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *9*(2), 930–940. https://doi.org/10.35957/jatisi.v9i2.1670

Kusumawardani, H. H., Rosyadi, I., Artanto, F. A., Arzha, F. I., & Rachmayani, N. A. (2022). Analisis decision tree dalam pengaruh digital marketing terhadap penerimaan siswa baru. *Remik: Riset Dan E-Jurnal Manajemen Informatika Komputer*, *6*(2), 225–231. http://doi.org/10.33395/remik.v6i2.11494

Mao, R., Shi, X., & Shi, Z. (2025). A Decision Tree Classification Algorithm Based on Two-Term RS-Entropy. *Entropy*, *27*(10). https://doi.org/10.3390/e27101069

Meng, L., Bai, B., Zhang, W., Liu, L., & Zhang, C. (2023). Research on a Decision Tree Classification Algorithm Based on Granular Matrices. *Electronics (Switzerland)*, *12*(21), 1–14. https://doi.org/10.3390/electronics12214470

Mienye, I. D., & Jere, N. (2024). A Survey of Decision Trees: Concepts, Algorithms, and Applications. *IEEE Access*, *12*(June), 86716–86727. https://doi.org/10.1109/ACCESS.2024.3416838

Nugraha, A. B., & Romadhony, A. (2023). Identification of 10 Regional Indonesian Languages Using Machine Learning. *Sinkron*, *8*(4), 2203–2214. https://doi.org/10.33395/sinkron.v8i4.12989

Nurain Jalaludin, Rahma Do Subuh, Ismail Maulud, & Bakhtiar Haerullah. (2024). Language Vitality Indicator Towards The Use Of Ternate Language On Ternate Language Enclave In Sahu District, West Halmahera Regency. *IJOLEH : International Journal of Education and Humanities*, *3*(1), 15–25. https://doi.org/10.56314/ijoleh.v3i1.209

Patasik, E. S., & Yulianto, S. (2023). Classification of Regional Languages Using Methods Gradient Boots and Random Forest. *Jurnal Teknik Informatika (Jutif)*, *4*(5), 1249–1255. https://doi.org/10.52436/1.jutif.2023.4.5.1459

Ramadhon, R. N., Ogi, A., Agung, A. P., Putra, R., Febrihartina, S. S., & Firdaus, U. (2024). Implementasi Algoritma Decision Tree untuk Klasifikasi Pelanggan Aktif atau Tidak Aktif pada Data Bank. *Karimah Tauhid*, *3*(2), 1860–1874. https://doi.org/10.30997/karimahtauhid.v3i2.11952

Saifudin, Fadlilah, N. I., Nouvel, A., & Sunanto, S. (2025). Penerapan Algoritma Decision Tree Dengan Optimasi Parameter Dalam Memprediksi Reaksi Autoimun Akibat Obat. *Informatics and Computer Engineering Journal*, *5*(2), 81–85. https://doi.org/10.31294/icej.v5i2.9157

Septiani, B., Hasanuddin, T., & Astuti, W. (2025). Classification of Lontara Script Using K-NN Algorithm, Decision Tree, and Random Forest Based on Hu Moments and Canny Segmentation. *Indonesian Journal of Data and Science*, *6*(2), 163–174. https://doi.org/10.56705/ijodas.v6i2.281

Sharma, D. N., & Iqbal, S. I. M. (2023). Applying Decision Tree Algorithm Classification and Regression Tree (CART) Algorithm to Gini Techniques Binary Splits. *International Journal of Engineering and Advanced Technology*, *12*(5), 77–81. https://doi.org/10.35940/ijeat.e4195.0612523

Shoaib, L. A., Safii, S. H., Idris, N., Hussin, R., & Sazali, M. A. H. (2024). Utilizing decision tree machine model to map dental students' preferred learning styles with suitable instructional strategies. *BMC Medical Education*, *24*(1), 1–13. https://doi.org/10.1186/s12909-023-05022-5

Sri Septiana, D. F., Triyanto, W. A., & Muzid, S. (2025). Penerapan Metode Decision Tree Algoritma C4.5 Dalam Sistem Rekomendasi Jurusan Bagi Calon Mahasiswa Baru. *Jurnal Unitek*, *18*(1), 167–178. https://doi.org/10.52072/unitek.v18i1.1322

Sujaini, H., & Putra, A. B. (2024). Analysis of language identification algorithms for regional Indonesian languages. *IAES International Journal of Artificial Intelligence*, *13*(2), 1741–1752. https://doi.org/10.11591/ijai.v13.i2.pp1741-1752

Surya, R., Dar, M. H., & Aini Nasution, F. (2024). Implementation of Decision Tree Method to Predict Customer Interest in Internet Data Packages. *International Journal of Science, Technology & Management*, *5*(4), 947–952. https://doi.org/10.46729/ijstm.v5i4.1155

Syatriawan, A., Fadlisyah, & Kurniawati. (2025). Penerapan Metode Decision Tree Cart Untuk Klasifikasi Penyakit Pada Tanaman Kelapa Sawit. *Rabit : Jurnal Teknologi Dan Sistem Informasi Univrab*, *10*(2), 1191–1199. https://doi.org/10.36341/rabit.v10i2.6544

Wahyudi, I., Annisa, A., Nst, S. A. E., & Manurung, A. S. (2025). The Role of Regional Languages as Markers of Cultural Identity. *HORIZON: Indonesian Journal of Multidisciplinary*, *3*(1), 38–45. https://doi.org/10.54373/hijm.v3i1.2427

Zaky, U., Naswin, A., Sumiyatun, S., & Murdiyanto, A. W. (2023). Performance Analysis of the Decision Tree Classification Algorithm on the Water Quality and Potability Dataset. *Indonesian Journal of Data and Science*, *4*(3), 145–150. https://doi.org/10.56705/ijodas.v4i3.113